HamleDT 3.0 License Terms

HamleDT 3.0 (referred to as "HamleDT" in the rest of this document) is a collection of linguistic textual data in multiple languages. It is based on pre-existing data sets ("original treebanks"). Each of the original treebanks has its own license terms and you (the "User") are responsible for complying with the license terms applicable to those parts of HamleDT, which you use. If you do not agree with the license terms, you must stop using HamleDT and destroy all copies of HamleDT data that you have obtained.

You are specifically reminded that some of the original treebanks permit only non-commercial usage.

The additional tree transformations and the software performing the transformations is copyright Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics ("we" or the "Provider"). It is provided as-is (without any warranty) and may be freely used, modified and redistributed under the GNU General Public License version 3.0 or the Perl Artistic License version 1.0.

The treebanks in HamleDT are organized in two groups: "Free" and "Patch". The Free group consists of treebanks whose license terms permit us redistributing them in full. For treebanks in the Patch group, we provide only our transformed annotation, without the underlying texts, lemmas and part of speech tags. If you legally obtained an original treebank, you can use the corresponding patch to transform the treebank to its HamleDT form.

Overview of the "Free" treebanks and their license terms

Note: the "-ud11" treebanks are taken from the Universal Dependencies release 1.1, see http://universaldependencies.github.io/docs/.

Code	Treebank	Web	License
ar	Prague Arabic Dependency Treebank	http://ufal.mff.cuni.cz/padt	CC BY-NC-SA 3.0
bg-ud11	BulTreeBank / UD1.1	http://www.bultreebank.org/indexBTB.html	CC BY-NC-SA 3.0
cs	Prague Dependency Treebank	http://ufal.mff.cuni.cz/pdt3.0	CC BY-NC-SA 3.0
da-ud11	Danish Dependency Treebank / UD1.1	http://www.buch-kromann.dk/matthias/treebank/	GNU GPL 2
de-ud11	UD1.1 (from Google)	http://universaldependencies.github.io/docs/#language-de	CC BY-NC-SA 3.0 US
el-ud11	Greek Dependency Treebank / UD1.1	http://universaldependencies.github.io/docs/#language-el	CC BY-NC-SA 3.0
en-ud11	English Web Treebank (Stanford) / UD1.1	http://nlp.stanford.edu/software/stanford-dependencies.shtml	CC BY-SA 4.0
es-ud11	UD1.1 (from Google)	http://universaldependencies.github.io/docs/#language-es	CC BY-NC-SA 3.0 US
et	Eesti keele puudepank	http://vvv.cs.ut.ee/~kaili/Korpus/puud/	free download
eu-ud11	Basque Dependency Treebank / UD1.1	http://universaldependencies.github.io/docs/#language-eu	CC BY-NC-SA 3.0
fa	Persian Dependency Treebank	http://dadegan.ir/en/perdt	GNU GPL 3*

Code	Treebank	Web	License
fa-ud11	Uppsala Persian Dependency Treebank / UD1.1	http://universaldependencies.github.io/docs/#language-fa	CC BY-SA 4.0
fi-ud11	Turku Dependency Treebank / UD1.1	http://bionlp.utu.fi/fintreebank.html	CC BY-SA 4.0
fi-ud11ftb	FinnTreeBank / UD1.1	http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/	LGPLv3+ or CC BY 4.0 (dual license)
fr-ud11	UD1.1 (from Google)	http://universaldependencies.github.io/docs/#language-fr	CC BY-NC-SA 3.0 US
ga-ud11	Irish Dependency Treebank / UD1.1	http://universaldependencies.github.io/docs/#language-ga	CC BY-SA 3.0
grc	Ancient Greek Dependency Treebank	http://nlp.perseus.tufts.edu/syntax/treebank/	CC BY-NC-SA 2.5
he-ud11	Hebrew Dependency Treebank / UD1.1	http://www.cs.bgu.ac.il/~yoavg/data/hebdeptb/	CC BY-NC-SA 4.0
hr-ud11	SETimes.HR / UD1.1	http://universaldependencies.github.io/docs/#language-hr	CC BY-SA 4.0
hu-ud11	Szeged Treebank / UD1.1	http://www.inf.u-szeged.hu/projectdirs/hlt/hu/Treebank/treebank2.html	CC BY-NC-SA 3.0
id-ud11	UD1.1 (from Google)	http://universaldependencies.github.io/docs/#language-id	CC BY-NC-SA 3.0 US
it-ud11	Italian Stanford Dependency Treebank / UD1.1	http://universaldependencies.github.io/docs/#language-it	CC BY-NC-SA 3.0
la	Latin Dependency Treebank	http://nlp.perseus.tufts.edu/syntax/treebank/	CC BY-NC-SA 2.5
la-it	Index Thomisticus Treebank	http://itreebank.marginalia.it/	CC BY-NC-SA 3.0
nl	CoNLL 2006 (Alpino)	http://odur.let.rug.nl/~vannoord/trees/	GNU GPL
pl	Składnica zależnościowa 0.5	http://zil.ipipan.waw.pl/Sk%C5%82adnica	GNU GPL 3
pt	CoNLL 2006 (Floresta Sintá(c)tica)	http://www.linguateca.pt/Floresta/principal.html	CC BY-NC-SA 3.0
ro	Resurse pentru Gramaticile de Dependenta	http://www.phobos.ro/roric/texts/indexro.html	free download
sl	Slovene Dependency Treebank / CoNLL 2006	http://nl.ijs.si/sdt/	research, cite
sv-ud11	Talbanken05 / UD1.1	http://stp.lingfil.uu.se/~nivre/research/Talbanken05.html	CC BY-SA 4.0
ta	Tamil Dependency Treebank	http://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1/	CC BY-NC-SA 3.0

Overview of the “Patch” treebanks

Lang.	Treebank	Web
bn	Hyderabad Dependency Treebank / ICON 2010	
ca	AnCora-CA	http://clic.ub.edu/corpus/
de	TIGER Corpus / CoNLL 2009	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
en	Penn Treebank / CoNLL 2007	http://www.cis.upenn.edu/~treebank/
es	AnCora-ES	http://clic.ub.edu/corpus/
hi	Hyderabad Dependency Treebank / COLING 2012	
ja	Tübingen Treebank of Spoken Japanese (Tüba-J/S)	http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-js.html
ru	SynTagRus	http://www.ruscorpora.ru/en/search-syntax.html
sk	Slovak Treebank	http://korpus.sk/
te	Hyderabad Dependency Treebank / ICON 2010	
tr	METU-Sabanci (ODTÜ-Sabancı) Treebank	http://ii.metu.edu.tr/corpus

Licenses

CC BY-SA 4.0 http://creativecommons.org/licenses/by-sa/4.0/

CC BY-NC-SA 4.0 http://creativecommons.org/licenses/by-nc-sa/4.0/

CC BY-SA 3.0 http://creativecommons.org/licenses/by-sa/3.0/

CC BY-NC-SA 3.0 http://creativecommons.org/licenses/by-nc-sa/3.0/

CC BY-NC-SA 2.5 http://creativecommons.org/licenses/by-nc-sa/2.5/

GNU (L)GPL http://www.gnu.org/licenses/gpl.html

fa: Persian Dependency Treebank: The download page contained the statement “I will use the treebank for research purposes only.” The “Readme and Licence.txt” file says “only non-commercially”. However, the included license is the standard GPL 3 (without any restrictions).

References

The following research papers should be cited to give proper credit to the creators of the original treebanks and to us, the creators of HamleDT.

HamleDT: Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, Jan Hajič (2014): *HamleDT: Harmonized Multi-Language Dependency Treebank*. In: Language Resources and Evaluation, 48(4) pp. 601–637, Springer, Dordrecht, Netherlands, ISSN 1574-020X.

HamleDT 2.0: Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský (2014): HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of*

the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2334–2341. ELDA, Reykjavík, Iceland.

ar: Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, Petr Zemánek (2008): Prague Arabic Dependency Treebank: A Word on the Million Words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pp. 16–23. ELDA, Marrakech, Morocco.

bg: Kiril Simov, Petya Osenova (2005): Extending the Annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pp. 173–184. Barcelona, Spain.

ca, es: Mariona Taulé, Maria Antònia Martí, Marta Recasens (2008): AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. ELDA, Marrakech, Morocco.

cs: Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, Šárka Zikánová (2013): *Prague Dependency Treebank 3.0*. **http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3**. Charles University in Prague, ÚFAL, Praha, Czechia.

da: Matthias T. Kromann, Line Mikkelsen, Stine Kern Lynge (2004): *Danish Dependency Treebank*. http://code.google.com/p/copenhagen-dependency-treebank/. København, Denmark.

de: Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, George Smith (2002): The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

de-ud11, es-ud11, fr-ud11, id-ud11: Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, Jungmee Lee (2013): Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*. Sofija, Bulgaria.

el: Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, Stelios Piperidis (2005): Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pp. 149–160. Barcelona, Spain.

en: Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz (1993): Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics* 19:2, pp. 313–330.

en-ud11: Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, Christopher D. Manning (2014): A Gold Standard Dependency Corpus for English. In *Proceedings of LREC 2014*. Reykjavík, Iceland.

et: Eckhard Bick, Heli Uibo, Kaili Müürisep (2004): Arborest – a VISL-Style Treebank Derived from an Estonian Constraint Grammar Corpus. In *Proceedings of Treebanks and Linguistic Theories (TLT 2004)*.

eu: Itzair Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Aitzpea Garmendia, Maite Oronoz (2003): Construction of a Basque Dependency Treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*.

fa: Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, Behrouz Minaei-Bidgoli (2011): A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. In *5th Language and Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 227–231. Poznań, Poland.

fa-ud11: Mojgan Seraji (2015): Morphosyntactic Corpora and Tools for Persian (PhD thesis). *Studia Linguistica Upsaliensia 16*. Uppsala, Sweden.

fi: Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, Tapio Salakoski (2010): Treebanking Finnish. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pp. 79–90.

fi-ud11ftb: Atro Voutilainen, Tanja Purtonen, Kristiina Muhonen, Krister Lindén (2012): Specifying Treebanks, Outsourcing Parsebanks: FinnTreeBank 3. In *Proceedings of LREC 2012*. İstanbul, Turkey.

ga: Teresa Lynn, Jennifer Foster, Mark Dras, Lamia Tounsi (2014): Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study. In *Proceedings of CLTW 2014*. Dublin, Ireland.

grc, la: David Bamman, Gregory Crane (2011): The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage*, pp. 79–98. ISBN 978-3-642-20227-8. Springer, Berlin / Heidelberg, Germany.

he: Yoav Goldberg (2011): *Automatic Syntactic Processing of Modern Hebrew* (PhD thesis). Ben Gurion University, Israel.

hi, bn, te: Samar Husain, Prashanth Mannem, Bharat Ambati, Phani Gadde (2010): The ICON-2010 Tools Contest on Indian Language Dependency Parsing. In *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing*. Kharagpur, India.

hr: Željko Agić, Nikola Ljubešić (2014): The SETimes.HR Linguistically Annotated Corpus of Croatian. In *Proceedings of LREC 2014*, pp. 1724–1727. Reykjavík, Iceland.

hu: Dóra Csendes, János Csirik, Tibor Gyimóthy, András Kocsor (2005): The Szeged Treebank. In *Text, Speech and Dialogue (TSD)*, pp. 123–131. Springer, Berlin / Heidelberg, Germany.

it: Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte (2003): Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé (ed.): *Building and Using Parsed Corpora*, pp. 189–210. Kluwer, Dordrecht, Netherlands.

it-ud11: Cristina Bosco, Simonetta Montemagni, Maria Simi (2013): Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse (LAW VII & ID at ACL-2013)*, pp. 61–69. Sofija, Bulgaria.

ja: Yasuhiro Kawata, Julia Bartels (2000): *Stylebook for the Japanese Treebank in Verbmobil, Report 240*. Universität Tübingen, Tübingen, Germany.

la-it: Marco Passarotti, Felice Dell'Orletta (2010): Improvements in parsing the index thomisticus treebank. Revision, combination and a feature model for medieval Latin. In *Training* vol. 2, 61–024.

nl: Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, Begoña Villada (2002): Chapter 5. The Alpino Dependency Treebank. In *Algorithms for Linguistic Processing NWO PIONIER Progress Report*. Groningen, Netherlands.

pl: Alina Wróblewska, Adam Przepiórkowski (2014): Projection-based Annotation of a Polish Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC*, pp. 2306–2312. Reykjavík, Iceland.

pt: Susana Afonso, Eckhard Bick, Renato Haber, Diana Santos (2002): "Floresta sintá(c)tica": A Treebank for Portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. ELDA, Las Palmas, Spain.

ro: Mihaela Călăcean (2008): *Data-driven Dependency Parsing for Romanian*. Uppsala Universitet, Uppsala, Sweden.

ru: Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, Nadezhda Frid (2000): Dependency Treebank for Russian: Concept, Tools, Types of Information. In *Proceedings of the 18th Conference on Computational Linguistics, vol. 2*, pp. 987–991. ACL, Morristown, NJ, USA.

sk: Mária Šimková, Radovan Garabík (2006): Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006*, pp. 389–394. ISBN 5-288-04181-4. St. Petersburg University Press, Sankt-Peterburg, Russia.

sl: Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdeněk Žabokrtský, Andreja Žele (2006): Towards a Slovene Dependency Treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 1388–1391. ELDA, Genova, Italy.

sv: Joakim Nivre, Jens Nilsson, Johan Hall (2006): Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. ELDA, Genova, Italy.

ta: Loganathan Ramasamy, Zdeněk Žabokrtský (2012): Prague Dependency Style Treebank for Tamil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. ELDA, İstanbul, Turkey.

tr: Nart B. Atalay, Kemal Oflazer, Bilge Say (2003): The Annotation Process in the Turkish Treebank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC)*. EACL, Budapest, Hungary.